Optimal Private Halfspace Counting via Discrepancy

S. Muthukrishnan[*] Aleksandar Nikolov[†]

March 1, 2013

Abstract

A *range counting* problem is specified by a set P of size $|P| = n$ of points in $\mathbb{R}^d$, an integer *weight* x_p associated to each point $p \in P$, and a *range space* $\mathcal{R} \subseteq 2^P$. Given a query range $R \in \mathcal{R}$, the output is $R(\mathbf{x}) = \sum_{p \in R} x_p$. The *average squared error* of an algorithm $\mathcal{A}$ is $\frac{1}{|\mathcal{R}|} \sum_{R \in \mathcal{R}} (\mathcal{A}(R, \mathbf{x}) - R(\mathbf{x}))^2$. Range counting for different range spaces is a central problem in Computational Geometry.

We study (ε, δ)-differentially private algorithms for range counting. Our main results are for the range space given by hyperplanes, that is, the halfspace counting problem. We present an (ε, δ)-differentially private algorithm for halfspace counting in d dimensions which is $O(n^{1-1/d})$ approximate for average squared error. This contrasts with the $\Omega(n)$ lower bound established by the classical result of Dinur and Nissim [12] on approximation for arbitrary subset counting queries. We also show a matching lower bound of $\Omega(n^{1-1/d})$ approximation for any (ε, δ)-differentially private algorithm for halfspace counting.

Both bounds are obtained using discrepancy theory. For the lower bound, we use a modified discrepancy measure and bound approximation of (ε, δ)-differentially private algorithms for range counting queries in terms of this discrepancy. We also relate the modified discrepancy measure to classical combinatorial discrepancy, which allows us to exploit known discrepancy lower bounds. This approach also yields a lower bound of $\Omega((\log n)^{d-1})$ for (ε, δ)-differentially private *orthogonal* range counting in d dimensions, the first known superconstant lower bound for this problem. For the upper bound, we use an approach inspired by partial coloring methods for proving discrepancy upper bounds, and obtain (ε, δ)-differentially private algorithms for range counting with polynomially bounded shatter function range spaces.

1 Introduction

A *range counting* problem is specified by a set P of size $|P| = n$, and a *range space* $\mathcal{R} \subseteq 2^P$. Given a query range $R \in \mathcal{R}$, the output is $|\{p \in P \cap R\}|$. More generally, each point $p \in P$ has an integer *weight* x_p and the range returns $R(\mathbf{x}) = \sum_{p \in R} x_p$. This problem is fundamental in Computational Geometry and a workhorse in applications, for various examples of range spaces from axis-parallel boxes (orthogonal range counting), to regions bounded by hyperplanes (halfspace counting) and beyond (e.g., simplices). Orthogonal range counting is commonly used in databases and data analysis. Halfspace counting is not only interesting in itself, but general algebraic range counting can be "lifted" to a higher dimension and encoded as halfspace counting [33].

We study privacy of range counting. In private range counting the set P of points as well as the range space $\mathcal{R}$ are considered public information, while the point weights x_p are considered private (and may denote, e.g. number of users at a geographic location). As the exact solution can reveal the private weights, we need to turn to approximate solutions. We define the *average squared error* of an algorithm $\mathcal{A}$ for range counting as $\frac{1}{|\mathcal{R}|} \sum_{R \in \mathcal{R}} (\mathcal{A}(R, \mathbf{x}) - R(\mathbf{x}))^2$. For privacy, we adopt the well-established notion of differential privacy. A mechanism $\mathcal{M} = \{M_n\}$ is (ε, δ)-*differentially private* if for every n, every $\mathbf{x}, \mathbf{x}'$ with $\|\mathbf{x} - \mathbf{x}\|_1 \leq 1$, and every measurable $S \subseteq \mathbb{R}^d$, the map M_n satisfies

$$\Pr[M_n(\mathbf{x}) \in S] \leq e^{\varepsilon} \Pr[M_n(\mathbf{x}') \in S] + \delta. \quad (1)$$

[*] Rutgers University, muthu@cs.rutgers.edu

[†] Rutgers University anikolov@cs.rutgers.edu

Surprisingly, very little is known about private range counting. Applying methods of differential privacy from first principles (Laplace noise and the basic composition theorem of differential privacy) will add large — variance $\Omega(n^2)$ in the case of halfspace counting in the plane — noise to each output. More generally, let $\mathbf{A}$ be an incidence matrix for a range space $\mathcal{R}$ (i.e. a matrix whose rows are the indicator vectors of all ranges $R \in \mathcal{R}$) and let $\mathbf{x}$ be the weights. The problem of computing $\mathbf{Ax}$ is the range counting problem. The average squared error of an approximate algorithm $\mathcal{A}$ is $\frac{1}{|\mathcal{R}|}\|\mathcal{A}(\mathbf{x})-\mathbf{Ax}\|_2^2$. In general, we can consider this problem for any $\mathbf{A} \in \{0,1\}^{m\times n}$, not necessarily ones that correspond to natural ranges from some constant dimensional geometric space. This is the *predicate counting* problem, well-studied in differential privacy. Then it is known that no mechanism that has average squared error $o(n)$ can be (ϵ, δ)-differentially private [12, 15]. However, the lower bounds are obtained using random $\mathbf{A}$'s that will not correspond to specific range spaces of interest. No super-constant lower bounds are known against (ε, δ)-differential privacy for natural problems like halfspace or orthogonal range counting in constant dimensional space. [1]

Our results are for (ε, δ)-differentially private range counting, and use the combinatorial structure of $\mathbf{A}$'s for range spaces. Our main application is halfspace counting, but our approach is general and yields other results too.

- *(Halfspace counting upper bound)* The (primal) *shatter function* of $\mathcal{R}$ is defined as $\pi_{\mathcal{R}}(s) = \max_{X\in\binom{P}{s}} |\mathcal{R}|_X|$ (i.e. the number of *distinct* sets in the restriction $\mathcal{R}|_X$). The shatter function of $\mathcal{R}$ defined by halfspaces in d-dimensions is bounded as $\pi_{\mathcal{R}}(s) = O(s^d)$.

We show that there is an (ε, δ)-differentially private range counting mechanism that achieves $O(n^{1-1/d})$ average squared error for range spaces with shatter function bounded by $O(s^d)$, and therefore for d-dimensional halfspace range counting.

Our upper bound shows that previous lower bounds [12, 15] for general $\mathbf{A}$'s indeed do *not* apply to halfspace range counting. Our algorithm runs in time polynomial in n and m. Previous work on this problem is incomparable. Work by Blum, Ligett and Roth [4] gave a non-constructive squared error upper bound of $O(d^2 n^{4/3})$ for range spaces with VC-dimension d and a matching constructive bound for halfspace range counting for $(\varepsilon, 0)$-differential privacy with a slightly different objective. Since the shatter function of a range space with VC-dimension d is bounded by $O(s^d)$, our result also implies a constructive approximation upper bound of $O(n^{1-1/d})$ for VC-dimension d range spaces.

Our approach relies on prior work [25] to decompose the range space into a logarithmic number of range spaces, some of them consisting only of small ranges, and some containing a small number of distinct ranges. We exploit this trade-off between maximum range size and number of distinct ranges by combining randomized response and Laplacian noise based differentially private mechanisms, but this balancing still leaves us with large noise in some cases. Nevertheless, we can bound the average privacy loss over the points $p \in P$. Our main idea is to use this approach to preserve privacy for *most* points $p \in P$; the shatter function bound does not increase for restrictions of P and $\mathcal{R}$ and we can recurse on the remaining points of P. This argument is inspired by partial coloring methods used in discrepancy theory. ■

- *(Range counting lower bound)* For halfspace counting in d dimensions, we show that any mechanism that has average squared error within $o(n^{1-1/d})$ is not (ε, δ)-differentially private for any constant ε and δ. We prove this lower bound using a notion of *discrepancy* where, in contrast to the standard notion where $\{+1, -1\}$ colorings are considered, we allow $\{0, +1, -1\}$ colorings but subject to some budget constraints on $\{+1, -1\}$. The budget constraints allows us to relate this notion of discrepancy to the classical one. Once the approach via the correct notion of discrepancy is developed, the mechanics are simple. Lower bounds will follow from combinatorial analysis of the discrepancy of range spaces. For orthogonal range counting, our approach immediately gives a lower bound of $(\log n)^{d-O(1)}$ on the average squared error of any (ε, δ) differentially private mechanism. The best upper bound in this setting is the work of Chan, Shi, and Song [5] who give an algorithm with average squared error $O((\log n)^{2d})$. No previous super-constant lower bounds are known for this problem even for large constant d. We note that proving a tight lower bound on the combinatorial discrepancy of axis-aligned boxes in d dimensions is a major open problem in discrepancy theory, and any improvement to the current discrepancy lower bound will yield a corresponding improvement in lower bounds for privacy. ■

[1] Constant lower bounds follow from the work of Roth [27] as well as from reductions from lower bounds for conjunction queries.

In Section 2 we review related prior work. In Section 3, we define concepts we need, including differential privacy and suitable notions of discrepancy. In Section 4, we present our lower bounds, and in Section 5, the upper bounds. We describe extensions and alternative algorithmic solutions in Section 6.

2 Prior Work

There is a rich and growing literature on solving counting problems while satisfying strong privacy guarantees. We will survey the prior work that is most relevant to our results.

In a seminal paper, Dinur and Nissim [12] initiated the study of the limits of output perturbation in answering arbitrary counting queries privately. They showed that if an algorithm $\mathcal{A}$ satisfies $\|\mathcal{A}(\mathbf{x})-\mathbf{A}\mathbf{x}\|_\infty^2 = o(n)$ for a random 0-1 matrix $\mathbf{A}$, then an adversary can reconstruct $\mathbf{x}$ almost exactly, implying that the algorithm is not (ε, δ)-differentially private for any constant ε, δ.[2] There is relatively little prior work on negative results for (ϵ, δ)-differential privacy for natural restrictions of $\mathbf{A}$. An exception is the work on lower bounding the noise necessary to privately answer conjunction queries [24, 11]. Conjunction queries on a database with d attributes can be reduced to answering orthogonal range counting or halfspace range counting queries in d dimensions. When d is constant, the lower bounds on conjunction queries imply a lower bound of C^d (for an absolute constant $C > 1$) on the average squared error neccessary to answer d-dimensional halfspace or orthogonal queries privately (here and in the remainder of this section we suppress dependence on ε, δ, and the probability of failure). In other related work, Roth [27] showed that linear queries with fat shattering dimension D require squared noise $\Omega(D^2)$ to preserve privacy. The fat shattering dimension reduces to the VC-dimension for counting queries, and has value $d+1$ for the range space of halfspaces in d dimensions. No super-constant lower bounds were previously known for (ε, δ)-differential privacy for the halfspace range counting or orthogonal range counting problems in constant dimensional space.

The study of private range counting for restricted range spaces was initiated with the work of Blum, Ligett, and Roth [4], who, using an argument based on epsilon nets, showed that queries of VC dimension d can be answered with worst-case squared noise $O(d^2 n^{4/3})$. Their algorithm is not computationally efficient, but they gave efficient algorithms with comparable guarantees for the interval range counting and halfspace range counting problems. Although their error bound is inferior to ours (when the size of the database is comparable to the universe size), the models are not directly comparable. While we consider a finite universe, they consider a continuous space, but give relaxed utility guarantes, namely that each query answer is accurate for a halfspace close to the query halfspace. Additionally, their algorithms satisfy the stronger notion of $(\varepsilon, 0)$-differential privacy and accomodate the regime where $\|\mathbf{x}\|_1$ is public and bounded by n and P is much larger.

For interval queries, the work of Blum, Ligett, and Roth was subsequently improved by Xiao, Wang, and Gehrke [32] (in the regime where database size and universe size are comparable), who gave a polylogarithmic noise upper bound via the wavelet transform. A related algorithm that achieves an average squared error upper bound of $O((\log n)^{3d})$ for d-dimensional orthogonal range counting was given by Chan, Shi, and Song [5]. We note that if we relax the privacy guarantee of Chan, Shi, and Song to (ε, δ)-differential privacy, their algorithm can be analyzed to provide average squared error $O(\log^{2d} n)$.

Much subsequent work has focused on answering m arbitrary queries efficiently with squared error linear in n and polylogarithmic in m [16, 28, 20, 21, 19]. A related line of work investigates the problem of answering conjunction queries with optimal error [18, 2].

Prior work for $(\varepsilon, 0)$-differential privacy. Stronger lower bounds can be shown when $\delta = 0$, and there are known separations between the cases $\delta = 0$ and $\delta > 0$, even when δ is superpolynomially small [11]. Hardt and Tulwar [22] gave a lower bound for linear queries based on geometric properties of the query matrix $\mathbf{A}$. De [11] simplified and extended their lower bound results. Blum, Ligett, and Roth [4] showed that no $(\varepsilon, 0)$-differentially private mechanism can answer interval queries with any nontrivial noise when the universe is continuous.

Discrepancy theory. For background in discrepancy theory we refer the reader to the books of Chazelle [7] and Matoušek [26]. Chazelle provides an overview of the applications of discrepancy theory to computer

[2] Our methods based on discrepancy allow us to re-prove the lower bound of Dinur and Nissim, as well as the version of Dwork and Yekhanin [17] that uses an explicit $\mathbf{A}$.

science, while Matoušek gives a survey of discrepancy theory results for geometric range spaces.

Geometric range counting. Geometric range counting and the closely related problems of range sums and range searching have a rich history in computational geometry. We refer the reader to the survey of Agarwal and Erickson [1] for background.

3 Preliminaries

We typeset vectors and matrices as $\mathbf{x}$, $\mathbf{A}$ and their elements as x_j, A_{ij}. We denote the i-th row of $\mathbf{A}$ as $\mathbf{A}_{i*}$ and the j-th column as $\mathbf{A}_{*j}$. Given a matrix $\mathbf{A}$, the function $\mathrm{col}(\mathbf{A})$ equals the number of columns of $\mathbf{A}$. For a matrix $\mathbf{A}$ with n columns, and a set $S \subseteq [n]$ we use $\mathbf{A}|_S$ to denote the submatrix of $\mathbf{A}$ consisting of the columns corresponding to elements of S (with duplicated rows removed). Similarly, for a range space $\mathcal{R}$ with incidence matrix $\mathbf{A}$, the range space $\mathcal{R}|_S$ is the one corresponding to the incidence matrix $\mathbf{A}|_S$. We denote the i-th standard basis vector $(0, \ldots, 0, 1, 0, \ldots, 0)^T$ (where 1 is in the i-th coordinate) as $\mathbf{e_i}$. For a set P we denote the collection of subsets of P of size s as $\binom{P}{s}$.

3.1 Range Counting

We will use the definitions for range counting, average squared error, orthogonal and hyperspace range counting, as well as the linear algebraic notation introduced in the Introduction. We also consider worst-case squared error, which for an algorithm $\mathcal{A}$ and a range space with incidence matrix $\mathbf{A}$ is $\|\mathcal{A}(\mathbf{x}) - \mathbf{A}\mathbf{x}\|_\infty^2 \geq \frac{1}{m}\|\mathcal{A}(\mathbf{x}) - \mathbf{A}\mathbf{x}\|_2^2$. We give all our lower bounds in average squared error and state our upper bounds in terms of both average and worst-case squared error.

The *VC-dimension* of a range space $\mathcal{R}$ is defined as the size of the largest set $X \subseteq P$ such that $\mathcal{R}|_X = 2^X$. The (primal) *shatter function* of $\mathcal{R}$ is defined as $\pi_{\mathcal{R}}(s) = \max_{X \in \binom{P}{s}} |\mathcal{R}|_X|$ (i.e. the number of *distinct* sets in $\mathcal{R}|_X$).

Fact 1 ([26])**.** *If the VC-dimension of $\mathcal{R}$ is d, then $\pi_{\mathcal{R}}(s) = O(s^d)$. Conversely, if $\pi_{\mathcal{R}}(s) = s^{O(1)}$ then the VC-dimension of $\mathcal{R}$ is constant.*

Fact 2 ([26])**.** *The VC-dimension of the range space $\mathcal{R}$ induced on P by all halfspaces in $\mathbb{R}^d$ is $d+1$. The shatter function of $\mathcal{R}$ is bounded as $\pi_{\mathcal{R}} = O(s^d)$.*

3.2 Differential Privacy

For any two sets $\mathcal{U}$ (*the universe*) and Y, a *mechanism* $\mathcal{M}$ over $\mathcal{U}$ with range Y is a family of maps $\{M_n\}$, $M_n : \mathcal{U}^n \to \rho(Y)$, where $\rho(Y)$ is the set of random variables that take values in Y. For the rest of this paper, we will focus on mechanisms over $\mathbb{Z}$ or over $\{0,1\}$, with range $\mathbb{R}^m$.

Definition 1. *A mechanism $\mathcal{M} = \{M_n\}$ over (a subset of) $\mathbb{Z}$ with range Y is (ε, δ)-*differentially private *if for every n, every $\mathbf{x}, \mathbf{x}'$ with $\|\mathbf{x} - \mathbf{x}'\|_1 \leq 1$, and every measurable $S \subseteq Y$, the map M_n satisfies*

$$\Pr[M_n(\mathbf{x}) \in S] \leq e^\varepsilon \Pr[M_n(\mathbf{x}') \in S] + \delta.$$

For lower bounds we use the following claim, which implies that being able to decode most of the input from the output contradicts differential privacy.

Lemma 1 ([11])**.** *Let $\mathcal{M} = \{M_n\}$ be a mechanism such that for some n there exists a (not necessarily efficient) algorithm $\mathcal{A}$ such that*

$$\forall \mathbf{x} \in \mathbb{Z}^n : \Pr[\|\mathcal{A}(M_n(\mathbf{x})) - \mathbf{x}\|_1 > \alpha n] < \beta.$$

Then there exist $\varepsilon = \varepsilon(\alpha, \beta)$ and $\delta = \delta(\alpha, \beta)$ such that the mechanism $\mathcal{M}$ is not (ε, δ)-differentially private.

A basic mechanism to achieve differential privacy with $\delta = 0$ is the *Laplace noise mechanism*, first proposed in [13]. Let us here and for the rest of the paper denote by $\mathrm{Lap}(s)$ the Laplace distribution centered at 0 with scale parameter s.

Lemma 2 ([13])**.** *Let f be any real-valued function which for any $\mathbf{x}, \mathbf{x}' \in \mathbb{Z}^n$ such that $\|\mathbf{x}-\mathbf{x}'\|_1 \leq 1$ satisfies $|f(\mathbf{x}) - f(\mathbf{x}')| \leq 1$. Then the mechanism that on input $\mathbf{x}$ outputs $f(\mathbf{x}) + \mathrm{Lap}(1/\varepsilon)$ satisfies $(\varepsilon, 0)$-differential privacy.*

The *composition* of mechanisms $\mathcal{M}^1 = \{M_n^1\}$, …, $\mathcal{M}^s = \{M_n^s\}$ is the mechanism that on input $\mathbb{Z}^n$ outputs $(M_n^1(\mathbf{x}), \ldots, M_n^s(\mathbf{x}))$. We need the following composition lemma first proved in [13].

Lemma 3 ([13])**.** *Let the mechanisms $\mathcal{M}^1, \ldots, \mathcal{M}^s$ satisfy, respectively, $(\varepsilon_1, \delta_1), \ldots, (\varepsilon_s, \delta_s)$ differential privacy. The composition $\mathcal{M}$ of the mechanisms satisfies $(\sum_i \varepsilon_i, \sum_i \delta_i)$-differential privacy.*

We also need a stronger result, which is a straightforward extension of the composition theorem of Dwork, Rothblum, and Vadhan [14]. To state the result we define a notion of privacy loss. Following [14], let us first define the *maximum divergence* of two random variables a and b as

$$D_\infty(a\|b) = \max_S \ln \frac{\Pr[a \in S]}{\Pr[b \in S]},$$

where S ranges over measurable subsets of the support of b. Note that a mechanism $\mathcal{M} = \{M_n\}$ is $(\varepsilon, 0)$-differentially private if and only if for every n and any $\mathbf{x}, \mathbf{x}' : \|\mathbf{x}-\mathbf{x}'\|_1 \leq 1$, we have $D_\infty(M_n(\mathbf{x})\|M_n(\mathbf{x}')) \leq \varepsilon$ and $D_\infty(M_n(\mathbf{x}')\|M_n(\mathbf{x})) \leq \varepsilon$.

Definition 2. *Let $\mathcal{M}$ be a composition of $\mathcal{M}^1, \ldots, \mathcal{M}^s$. The* privacy loss of $i \in [n]$ for the j-th output *is*

$$l_{\mathcal{M}}(i, j) = \max_{\mathbf{x}, \mathbf{x}' = \mathbf{x} \pm \mathbf{e_i}} D_\infty(M_n^j(\mathbf{x})\|M_n^j(\mathbf{x}')),$$

The (ℓ_2) privacy loss of $i \in [n]$ *is* $L_{\mathcal{M}}(i) = \sqrt{\sum_{j \in [s]} l_{\mathcal{M}}(i, j)^2}$.

Lemma 4. *Let $\mathcal{M}$ be a composition of $\mathcal{M}^1, \ldots, \mathcal{M}^s$ and let $\varepsilon > \max_{i \in [n]} L_{\mathcal{M}}(i)$. Then, for any $\delta > 0$, $\mathcal{M}$ satisfies $(\sqrt{2\ln(1/\delta)}\varepsilon, \delta)$-differential privacy.*

Note that for the range counting problem, the privacy loss is defined for a point p.

3.3 Discrepancy

Here we define a modified notion of discrepancy. In Section 4, we show that this modified notion of discrepancy is useful in carrying out Dinur-Nissm type attacks on privacy.

Definition 3. *For any $\mathbf{A} \in \mathbb{R}^{m \times n}$, we define*

$$\mathrm{disc}_{p,\alpha}(\mathbf{A}) = \min_{\substack{\mathbf{x} \in \{0, \pm 1\}^n \\ \|\mathbf{x}\|_1 \geq \alpha\,\mathrm{col}(A)}} \|\mathbf{A}\mathbf{x}\|_p$$
$$\mathrm{herdisc}_{p,\alpha}(\mathbf{A}) = \max_{S \subseteq [n]} \mathrm{disc}_{p,\alpha}(\mathbf{A}|_S).$$

The standard notions of discrepancy and hereditary discrepancy correspond to the special cases disc $=$ $\mathrm{disc}_{\infty,1}$ and herdisc $=$ $\mathrm{herdisc}_{\infty,1}$. The cases $\mathrm{disc}_{2,1}$ and $\mathrm{herdisc}_{2,1}$ have also been extensively studied, especially as means of proving lower bounds on disc and herdisc. On the other hand the case $\mathrm{disc}_{p,0}$ is trivially the identically 0 function. Next, we exhibit a connection between $\mathrm{herdisc}_{p,1}$ and $\mathrm{herdisc}_{p,\alpha}$ for $\alpha \in (0, 1)$ and any p.

Lemma 5. *Let $f(s) = \max_{S \subseteq [n]: |S| \leq s} \mathrm{disc}_{p,\alpha}(\mathbf{A}|_S)$. Then $\mathrm{disc}_{p,1}(\mathbf{A}) \leq \sum_{i=0}^\infty f((1-\alpha)^i n)$, and, therefore, $\mathrm{herdisc}_{p,1}(\mathbf{A}) \leq \sum_{i=0}^\infty f((1-\alpha)^i n)$*

Proof. We will find an assignment $\mathbf{x} \in \{\pm 1\}^n$ such that $\|\mathbf{A}\mathbf{x}\|_p \leq \sum_{i=0}^\infty f((1-\alpha)^i n)$, which is sufficient to prove the lemma. Let $\mathbf{x}' \in \{0, \pm 1\}^n$ be such that $\|\mathbf{A}\mathbf{x}\|_p \leq f(n)$ and $\|\mathbf{x}\|_1 \geq \alpha n$. Let $S = \{i : x_i = 0\}$. Since $\|\mathbf{x}\|_1 \geq \alpha n$, $|S| \leq (1-\alpha)n$. We recurse to find an assignment $\mathbf{x}'' \in \{\pm 1\}^S$ such that $\|(\mathbf{A}|_S)\mathbf{x}''\|_p \leq \sum_{i=0}^\infty f((1-\alpha)^i |S|) \leq \sum_{i=1}^\infty f((1-\alpha)^i n)$. Set $x_i = x_i'$ when $i \notin S$ and $x_i = x_i''$ when $i \in S$. □

Lemma 5 and the observation $\mathrm{herdisc}_{p,\alpha} = \max_{s=1}^{n} f(s)$ imply that for any $\mathbf{A}$,

$$\mathrm{herdisc}_{p,1}(\mathbf{A}) \leq \frac{\log n}{\log 1/(1-\alpha)} \mathrm{herdisc}_{p,\alpha}(\mathbf{A}).$$

However using Lemma 5 directly and the observation that a restriction of a halfspace range space (or a range space of axis-aligned boxes) is a range space of the same kind, we get stronger lowerbounds for $\mathrm{herdisc}_{p,\alpha}$. Below we list several interesting results that can be derived in this way from known results in combinatorial discrepancy theory [7, 26]. Below we provide more specific references to the discrepancy lower bound used to derive each result. We provide a full proof of the first result; the remaining proofs follow analogous reasoning.

Lemma 6 ([9]). *For infinitely many n there exists a set of n points P and m halfspaces $H_1, \ldots, H_m$ in $\mathbb{R}^d$ ($d = O(1)$) such that the following holds. Let $\mathbf{A}$ denote the incidence matrix of the collection of sets $\{H_j \cap P, j \in [m]\}$. Then for any $\alpha = \Omega(1)$, $\mathrm{herdisc}_{2,\alpha}(\mathbf{A}) = \Omega(m^{1/2} n^{1/2-1/2d})$.*

Proof. Assume for contradiction that all but finately many $m \times n$ incidence matrices $\mathbf{A}$ of halfspaces in $\mathbb{R}^d$ have hereditary α-discrepancy $\mathrm{herdisc}_{2,\alpha}(\mathbf{A}) = o(m^{1/2} n^{1/2-1/2d})$. By the results in [9], there exist infinitely many sets of n points P and $m = \binom{n}{d}$ halfspaces $H_1, \ldots, H_m$ such that the incidence matrix $\mathbf{B}$ of $\{H_j \cap P, j \in [m]\}$ has hereditary discrepancy $\mathrm{herdisc}_{2,1} = \Omega(m^{1/2} n^{1/2-1/2d})$. Let us fix any such set of points and halfspaces and the corresponding incidence matrix $\mathbf{B}$. Any restriction $\mathbf{B}|_S$ for $S \subseteq P$ is also the incidence matrix of sets induced by points and halfspaces, and by assumption, $\mathrm{herdisc}(\mathbf{B}|_S) = o(m^{1/2}|S|^{1/2-1/2d})$. Plugging this bound in Lemma 5 we get $\mathrm{herdisc}_{2,1}(\mathbf{B}) = o(m^{1/2} n^{1/2-1/2d})$, a contradiction. □

Lemma 7 ([29, 3]). *For infinitely many n there exists a set of n points P and m axis-parallel boxes $B_1, \ldots, B_m$ in $\mathbb{R}^d$ ($d = O(1)$) such that the following holds. Let $\mathbf{A}$ denote the incidence matrix of the collection of sets $\{B_j \cap P, j \in [m]\}$. Then for any $\alpha = \Omega(1)$, $\mathrm{herdisc}_{2,\alpha}(\mathbf{A}) = \Omega(m^{1/2} (\log n)^{d/2-3/2})$.*

Lemma 8 ([8]). *For infinitely many n there exists a set of n points P and m axis-parallel boxes $B_1, \ldots, B_m$ in $\mathbb{R}^d$ ($d = \Theta(\log n)$) such that the following holds. Let $\mathbf{A}$ denote the incidence matrix of the collection of sets $\{B_i \cap P, j \in [m]\}$. Then for any $\alpha = \Omega(1)$, $\mathrm{herdisc}_{\infty,\alpha}(\mathbf{A}) = n^{\Omega(1)}$.*

Lemma 9 ([30]). *For any n and $m > n$ there exists a matrix $\mathbf{A} \in \{0,1\}^{m \times n}$ such that $\mathrm{herdisc}_{\infty,\alpha}(A) = \Omega(\sqrt{n \log 2m/n})$.*

4 Lower Bounds for Privacy from Discrepancy

Our main result in this section is a noise lower bound on (ε, δ)-differentially private mechanisms that approximate range counting queries for a host of natural geometric range spaces. Our main conceptual contribution is in identifying $\mathrm{herdisc}_{p,\alpha}$ as the key quantity in showing lower bounds against (ε, δ)-differential privacy via a Dinur-Nissim type attack, and connecting this quantity to the standard notion of combinatorial discrepancy.

Theorem 1. *For any α, β, there exist $\varepsilon(\alpha, \beta)$ and $\delta(\alpha, \beta)$ such that no mechanism $\mathcal{M} = \{M_n\}$ over the universe $\{0,1\}$ with range $\mathbb{R}^m$ that for some p satisfies*

$$\forall \mathbf{x} \in \{0,1\}^n : \Pr[\|M_n(\mathbf{x}) - \mathbf{A}\mathbf{x}\|_p < \mathrm{disc}_{p,\alpha}(\mathbf{A})/2] \geq 1 - \beta,$$

is (ε, δ)-differentially private.

We extend the lower bound to $\mathrm{herdisc}_{p,\alpha}$. This allows us to use the connection between $\mathrm{herdisc}_{p,\alpha}$ and standard discrepancy.

Corollary 1. *For any α, β, there exist $\varepsilon(\alpha, \beta)$ and $\delta(\alpha, \beta)$ such that no mechanism $\mathcal{M} = \{M_n\}$ over the universe $\{0,1\}$ with range $\mathbb{R}^m$ that for some p satisfies*

$$\forall \mathbf{x} \in \{0,1\}^n : \Pr[\|M_n(\mathbf{x}) - \mathbf{A}\mathbf{x}\|_p < \mathrm{herdisc}_{p,\alpha}(\mathbf{A})/2] \geq 1 - \beta,$$

is (ε, δ)-differentially private.

Proof. We claim that given M_n and any set $S \subseteq [n]$, we can construct M_n' that takes as input $\mathbf{x}|_S$, is (ε, δ)-differentially private (with respect to $\mathbf{x}|_S$), and satisfies

$$\forall \mathbf{x}|_S : \Pr[\|M_n'(\mathbf{x}|_S) - (\mathbf{A}|_S)(\mathbf{x}|_S)\|_p < \mathrm{herdisc}_{p,\alpha}(\mathbf{A}|_S)/2] \geq 1 - \beta.$$

Then we can take S such that $\mathrm{disc}_{p,\alpha}(\mathbf{A}|_S) = \mathrm{herdisc}_{p,\alpha}(\mathbf{A})$, and the corollary follows from Theorem 1.

We define M_n' as follows: $M_n'(\mathbf{x}|_S)$ extends $\mathbf{x}|_S$ to $\mathbf{x}$ by setting $x_i = 0$ for all $i \notin S$ and outputs $M_n(\mathbf{x})$. It's easy to verify that M_n' satisfies the claimed properties. □

Theorem 1 follows from Lemma 1 and the following lemma.

Lemma 10. *There exists a deterministic (not necessarily efficient) algorithm $\mathcal{A}$ that on input a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and a vector $\tilde{\mathbf{y}} \in \mathbb{R}^m$ satisfying $\|\tilde{\mathbf{y}} - \mathbf{A}\mathbf{x}\|_p < \mathrm{disc}_{p,\alpha}(\mathbf{A})/2$ for some $\mathbf{x} \in \{0,1\}^n$, outputs a vector $\mathbf{x}' \in \{0,1\}^n$ such that $\|\mathbf{x}' - \mathbf{x}\|_1 \leq \alpha n$.*

Proof. Given $\tilde{\mathbf{y}}$, $\mathcal{A}$ outputs an arbitrary $\mathbf{x}' \in \{0,1\}^n$ such that $\|\mathbf{A}\mathbf{x}' - \tilde{\mathbf{y}}\|_p < \mathrm{disc}_{p,\alpha}(\mathbf{A})/2$. Such a $\mathbf{x}'$ exists, since $\|\mathbf{A}\mathbf{x} - \tilde{\mathbf{y}}\|_p < \mathrm{disc}_{p,\alpha}(\mathbf{A})/2$ by assumption. We claim that $\|\mathbf{x} - \mathbf{x}'\| \leq \alpha n$. For contradiction, assume $\|\mathbf{x} - \mathbf{x}'\| > \alpha n$. Notice that $\mathbf{x} - \mathbf{x}' \in \{0, \pm 1\}$. Then, by the definition of $\mathrm{disc}_{p,\alpha}$, $\|\mathbf{A}(\mathbf{x} - \mathbf{x}')\|_p \geq \mathrm{disc}_{p,\alpha}(\mathbf{A})$. By the triangle inequality, the assumption of the lemma, and the definition of $\mathcal{A}$, $\|\mathbf{A}(\mathbf{x} - \mathbf{x}')\|_p \leq \|\mathbf{A}\mathbf{x} - \tilde{\mathbf{y}}\|_p + \|\mathbf{A}\mathbf{x}' - \tilde{\mathbf{y}}\|_p < \mathrm{disc}_{p,\alpha}(\mathbf{A})$, and we've reached a contradiction. □

Corollary 1, instantiated with $p = 2$, and Lemmas 6–8 imply an array of noise lower bounds for approximating geometric range counting while satisfying (ε, δ)-differential privacy.

Theorem 2. *Any mechanism $\mathcal{M}$ that, for any P in $\mathbb{R}^d$ with $|P| = n$ and $d = O(1)$, with constant probability approximates the halfspace range counting problem within average squared error $o(n^{1-1/d})$ is not (ε, δ)-differentially private for any constant ε and δ.*

Theorem 3. *Any mechanism $\mathcal{M}$ that, for any P in $\mathbb{R}^d$ with $|P| = n$ and $d = O(1)$, with constant probability approximates the orthogonal range counting problem within average squared error $o((\log n)^{d-1})$ is not (ε, δ)-differentially private for any constant ε and δ.*

Theorem 4. *Any mechanism $\mathcal{M}$ that, for any P in $\mathbb{R}^d$ with $|P| = n$ and $d = \Theta(\log n)$, with constant probability approximates the orthogonal range counting problem within average squared error $n^{o(1)}$ is not (ε, δ)-differentially private for any constant ε and δ.*

We also note that that Corollary 1, instantiated with $p = \infty$ and Lemma 9 imply a lower bound on the worst case squared error for privately approximating m arbitrary range counting queries where m is much larger than n.

Theorem 5. *Any mechanism $\mathcal{M}$ that, for any range space $(P, \mathcal{R})$ ($|P| = n$, $|\mathcal{R}| = m$), with constant probability approximates range counts for $\mathcal{R}$ with worst case squared error $o(n \log 2m/n)$ is not (ε, δ)-differentially private for any constant ε and δ.*

The results of Dinur and Nissim [12] for $m = \emptyset(n)$ and $m = 2^n$ are special cases of Theorem 5. To the best of our knowledge, this is the first lower bound that explicitly accounts for the dependence of error on m for arbitrary $m > n$.

5 Algorithm for Bounded Shatter Function Systems

In this section we present an efficient (for constant d) (ε, δ)-differentially private range counting algorithm for range spaces with bounded shatter function. We prove the algorithm gives optimal average squared error and almost optimal worst-case squared error bounds. The algorithm is based on a novel use of a decomposition that was first constructed by Matoušek [25] to prove optimal discrepancy upper bounds for bounded shatter function range spaces. Even a careful application of known methods in differential privacy together with the decomposition does not provide optimal error bounds directly; we, however, prove that privacy can be satisfied for a constant fraction of P while achieving optimal error bounds; then we recurse on the remainder of P. Aside from the decomposition, this method of satisfying privacy for a fraction of the database is inspired by partial coloring methods in discrepancy theory.

We will make an essential use of the following lemma, due originally to Haussler. The lemma bounds the size of an epsilon net in the hamming metric.

Lemma 11 ([23])**.** *Let $(P, \mathcal{R})$ be a range space with shatter function $\pi_{\mathcal{R}}(s) = O(s^d)$. Let Δ be an integer less than $|P|$. Let $\mathcal{S} \subseteq \mathcal{R}$ be a collection of ranges such that for any two ranges $R_1, R_2 \in \mathcal{S}$, the symmetric difference between R_1 and R_2 is at least Δ. Then, $|\mathcal{S}| = O((|P|/\Delta)^d)$.*

We construct collections of ranges with large pairwise distance Δ for gemetrically growing values of Δ. Using the collections as finer and finer epsilon nets, we can represent each range in $\mathcal{R}$ as the union and set difference of smaller and smaller ranges, while Lemma 11 allows us to control the number of such ranges needed for each value of Δ. We then approximate range counts for the ranges that make up the decomposition; the trade-off between range size and number of distinct ranges allows us to balance the noise incurred by randomized response and by using composition (Lemma 4).

We first detail the construction. Our presentation follows [26]. Let $(P, \mathcal{R})$ be a range space with shatter function $\pi_{\mathcal{R}}(s) = O(s^d)$. Let $k = \lceil \log_2 n \rceil$. For each $i \in \{0, \dots, k\}$, let $\mathcal{S}_i \subseteq \mathcal{R}$ be a maximal collection of ranges such that the symmetric difference between any two ranges $R_1, R_2 \in \mathcal{S}_i$ is at least $n2^{-i}$. In particular, $\mathcal{S}_k = \mathcal{R}$ and $\mathcal{S}_0 = \{\emptyset\}$. For each $R \in \mathcal{S}_i$, fix a $R' \in \mathcal{S}_{i-1}$ such that the symmetric difference between R and R' is at most $n2^{-i+1}$ (such a range exists by maximality of $\mathcal{S}_{i-1}$). Then we set $F(R) = R \setminus R'$ and $G(R) = R' \setminus R$, so that $R' = (R \setminus F(R)) \cup G(R)$, $F(R) \subseteq R$, and $G(R) \cap (R \setminus F(R)) = \emptyset$. Define a new collection of ranges $\mathcal{T}_i = \{F(R), G(R) : R \in \mathcal{S}_i\}$. We can start from $R \in \mathcal{R} = \mathcal{S}_k$ and apply the construction recursively, until we have $\emptyset = ((\dots((R \setminus F_k) \cup G_k)\dots) \cup G_2) \setminus F_1$, where $F_i, G_i \in \mathcal{T}_i$. Bactracking to reconstruct R, we get

$$R = ((\dots(F_1 \setminus G_2) \cup F_2 \dots) \setminus G_k) \cup F_k. \tag{2}$$

All union operations are on disjoint sets and any set is subtracted from a set that entirely contains it.

Each range in $\mathcal{T}_i$ has size at most $n2^{-i+1}$ by construction; by Lemma 11, $|\mathcal{S}_i| = O(2^{di})$, and, since each range in $\mathcal{S}_i$ corresponds to at most two ranges in $\mathcal{T}_i$, we also have $\mathcal{T}_i = O(2^{di})$. Let $\mathbf{T}^i$ be the incidence matrix of $\mathcal{T}_i$. The following lemma follows from the decomposition (2):

Lemma 12. *Let $(P, \mathcal{R})$ be a range space with $|P| = n$ and shatter function $\pi_{\mathcal{R}}(s) = O(s^d)$. Let $\mathbf{A}$ be the incidence matrix of $\mathcal{R}$. Then, there exist matrices $\mathbf{T}^i \in \{0,1\}^{s_i \times n}$ and $\mathbf{Q}^i \in \{0, \pm 1\}^{m \times s_i}$ such that $\mathbf{A} = \sum_{i=1}^{k} \mathbf{Q}^i \mathbf{T}^i$. Furthermore, we have the following properties for $\mathbf{T}^i$ and $\mathbf{Q}^i$:*

- *each row in $\mathbf{T}^i$ has at most $n2^{-i+1}$ nonzero entries;*
- *$s_i \leq C2^{di}$ for some absolute constant C;*
- *each row in $\mathbf{Q}^i$ has at most 2 nonzero entries.*

For the degree of a point $p \in P$ in the range space $\mathcal{T}_i$, we use the notation $d_i(p) = |\{R \in \mathcal{T}_i : p \in R\}|$.

Intuitively, we will use randomized response on those $\mathcal{T}_i$ consisting of only small ranges, and we will use the Laplace noise mechanism on those $\mathcal{T}_i$ consisting of few ranges. The "breaking-even point" for the analysis is $i_0 = (\log n)/d$. For $i \geq i_0$ randomized response gives the guarantee we need: the largest range in $\mathcal{T}_i$ for $i \geq i_0$ has size at most $n^{1-1/d}$. However, $\mathcal{T}_{i_0}$ can have as many as n ranges, and it seems that we cannot use Laplace noise with variance $n^{1-1/d}$ and still preserve privacy for those i close to i_0. To circumvent this issue, we use the fact that we can bound both the largest range and the number of ranges in each $\mathcal{T}_i$ simultaneously. The main observation is that we can add noise with optimal variance $O(n^{1-1/d})$ to the range counts for those $\mathcal{T}_i$ where randomized response doesn't work, and bound the average privacy loss $\frac{1}{n}\sum_p L_{\mathcal{M}}(p)$. Then, we use averaging and Lemma 4, and argue that we can preserve privacy for *most* $p \in P$. The shatter function bound does not increase for restrictions of P and $\mathcal{R}$ and we can recurse on the remaining points of P. Our algorithm for computing range counts over ranges with bounded shatter function is given as Algorithm 1. The algorithm description and the following discussion assume that $\mathcal{R}$ has shatter function $\pi_{\mathcal{R}}(s) = O(s^d)$ (for $d \geq 2$) and the decomposition of Lemma 12 has already been computed. Note that the decomposition can be computed in time $O(mn \log n)$.

We analyze the privacy guarantees of Algorithm 1. We first prove some technical claims about the algorithm.

Algorithm 1 $\textsc{RangeCount}(P, \mathbf{x}, \mathcal{R}, \varepsilon, \delta)$

```
  Let $|P| = n$, $|\mathcal{R}| = m$;
  Set $i_0 := \frac{\log n}{d}$;
  Set $\varepsilon_i := \frac{\varepsilon(i-i_0+1)^{1.5}}{n^{1/2-1/2d}}$ for $i \le i_0$;
  Set $\varepsilon_i := \frac{\varepsilon}{(i-i_0+1)^{1.5}}$ for $i > i_0$;
  if $n \le 1$ then
    Let $p \in P$ be the only point in $P$. Return $\tilde{x}_p := x_p + \mathrm{Lap}(1/\varepsilon)$ for all $R \in \mathcal{R}$ s.t. $p \in R$ and 0 for all other $R \in \mathcal{R}$.
  end if

  Set $X := \{p : \sum_{i \le i_0} d_i(p)\varepsilon_i^2 \le 12C\varepsilon^2\}$ and $\bar{X} := P \setminus X$;
  Recursively compute $\textsc{RangeCount}(\bar{X}, \mathbf{x}|_{\bar{X}}, \mathcal{R}|_{\bar{X}}, \varepsilon, \delta)$; let the results be $\tilde{z}_1^1, \dots, \tilde{z}_m^1$.
  for all $i \le i_0$ do
    Compute $\tilde{\mathbf{y}}^i := (\mathbf{T}^i|_X)(\mathbf{x}|_X) + \mathrm{Lap}(1/\varepsilon_i)^{s_i}$;
  end for
  for all $i_0 < i \le k$ do
    Compute $\tilde{\mathbf{x}}^i := \mathbf{x} + \mathrm{Lap}(1/\varepsilon_i)^n$;
    Compute $\tilde{\mathbf{y}}^i := (\mathbf{T}^i|_X)(\tilde{\mathbf{x}}^i|_X)$;
  end for

  Compute $\tilde{\mathbf{z}}^2 := \sum_{i=1}^k \mathbf{Q}^i \tilde{\mathbf{y}}^i$;
  Output $\tilde{\mathbf{z}} = \tilde{\mathbf{z}}^1 + \tilde{\mathbf{z}}^2$.
```

Lemma 13. *The following hold for Algorithm 1:*

1. $|X| \ge n/2$.
2. $\{\tilde{\mathbf{y}}^i\}_{i=1}^{i_0}$ *is a* $(2\sqrt{6C}\varepsilon\sqrt{\ln(1/\delta)}, \delta)$*-differentially private function of* $\mathbf{x}|_X$.
3. $\{\tilde{\mathbf{x}}^i\}_{i=i_0+1}^k$ *is a* $(2\varepsilon, 0)$*-differentially private function of* $\mathbf{x}$. *Moreover, for each* $S \subseteq P$, $\{\tilde{\mathbf{x}}^i|_S\}_{i=i_0+1}^k$ *is a* $(2\varepsilon, 0)$*-differentially private function of* $\mathbf{x}|_S$.

Proof. Claim 1. follows by avaraging and the inequality

$$\frac{1}{n}\sum_{p \in P}\sum_{i < i_0} d_i(p)\varepsilon_i^2 \le 6C\varepsilon^2 \tag{3}$$

Next we establish (3).

$$\begin{aligned}\frac{1}{n}\sum_{p \in P}\sum_{i \le i_0} d_i(p)\varepsilon_i^2 &\le \frac{1}{n}\sum_{i \le i_0} C2^{di} n 2^{-i+1}\frac{\varepsilon^2(i-i_0+1)^3}{n^{1-1/d}} \\ &\le 2C\varepsilon^2 \sum_{j=0}^{\infty}\frac{(j+1)^3}{2^{dj+j}} \le 6C\varepsilon^2.\end{aligned}$$

The first inequality follows from Lemma 12. The second inequality holds for $d \ge 2$. This finishes the proof of claim 1.

The following privacy analysis uses the fact that the range space $(P, \mathcal{R})$ is public, and, therefore, the decomposition given by Lemma 12, and the set X determined by the decomposition are public as well, i.e. independent of $\mathbf{x}$.

Notice that each component of $\tilde{\mathbf{y}}^i$ is an instance of the Laplace noise mechanism and, therefore, by Lemma 2 it is $(\varepsilon_i, 0)$-differentially private. Also, $\tilde{y}_j^i$ is independent of x_p whenever $T_{jp}^i = 0$ or $p \notin X$. Denoting by $\tilde{y}_j^i(\mathbf{x})$ the random variable $\tilde{y}_j^i$ when the input is $\mathbf{x}$, we have that

$$D_\infty(\tilde{y}_j^i(\mathbf{x}) \| \tilde{y}_j^i(\mathbf{x} \pm \mathbf{e_p})) \le \begin{cases} 0, & T_{jp}^i = 0 \text{ or } p \notin X \\ \varepsilon_i, & \text{otherwise} \end{cases}$$

If $\mathcal{M}$ is the mechanism that outputs $\{\tilde{\mathbf{y}}^i\}_{i=1}^{i_0}$, then, by the above discussion, $L_{\mathcal{M}}(p) = \sqrt{\sum_{i \le i_0} d_i(p)\varepsilon_i^2}$. By the definition of X, we have that $L_{\mathcal{M}}(p) \le \sqrt{12C}\varepsilon$ for any $p \in X$ (and $L_{\mathcal{M}}(p) = 0$ for $p \notin X$). Claim 2. then follows by Lemma 4.

By Lemma 2, each $\tilde{\mathbf{x}}^i$ is $(\varepsilon_i, 0)$-differentially private. By Lemma 3, the composition $\{\tilde{\mathbf{x}}^i\}_{i=i_0+1}^k$ is $(\sum_{i=i_0+1}^k \varepsilon_i, 0)$-differentially private. Then claim 3. follows from

$$\sum_{i=i_0+1}^{k} \varepsilon_i < \varepsilon \sum_{j=2}^{\infty} j^{-1.5} < 2\varepsilon.$$

This completes the proof of the lemma. □

Theorem 6 (**Privacy**). *Algorithm 1 preserves* $((2\sqrt{6C}+2)\varepsilon\sqrt{\ln 1/\delta}, \delta)$*-differential privacy.*

Proof. We proceed by induction on n.

Base case. When $n \le 1$, the output of Algorithm 1 is $(\varepsilon, 0)$-differentially private, since it is a function of $\tilde{\mathbf{x}}$, which is itself $(\varepsilon, 0)$-differentially private by the properties of the Laplace noise mechanism (Lemma 2).

Inductive step. Note that $\tilde{\mathbf{z}}^2$ is a function of $\tilde{\mathbf{x}}|_X$ and $\{\tilde{\mathbf{y}}^i\}_{i=1}^{i_o}$. Also note that both $\tilde{\mathbf{x}}|_X$ and $\{\tilde{\mathbf{y}}^i\}_{i=1}^{i_o}$ depend only on X and not on $\bar{X}$. By simple composition (Lemma 3), and Lemma 13, $\tilde{\mathbf{z}}^2$ is a $((2\sqrt{6C}+2)\varepsilon\sqrt{\ln 1/\delta}, \delta)$-differentially private function of $\mathbf{x}|_X$. By Lemma 13, $\bar{X} < n/2$, so by the inductive hypothesis $\tilde{\mathbf{z}}^1$ is an $((2\sqrt{6C}+2)\varepsilon\sqrt{\ln 1/\delta}, \delta)$-differentially private function of $\mathbf{x}|_{\bar{X}}$. Since X and $\bar{X}$ are disjoint, it follows that $\tilde{\mathbf{z}} = \tilde{\mathbf{z}}^1 + \tilde{\mathbf{z}}^2$ is a $(6(\sqrt{C}+2)\sqrt{\ln 1/\delta}, \delta)$-differentially private function of $\mathbf{x}$. □

Next we analyze the approximation guarantee of the algorithm. The bounds in following lemma can derived by a straightforward calculation.

Lemma 14. *Let* $\mathbf{y}^i = (\mathbf{T}^i|_X)(\mathbf{x}|_X)$*. For each* $j \in [m]$ *and each* $i \le i_0$*,* $\mathbb{E}[\mathbf{Q}_{j*}^i\tilde{\mathbf{y}}^i] = \mathbf{Q}_{j*}^i\mathbf{y}^i$*, and* $\mathrm{Var}[\mathbf{Q}_{j*}^i\tilde{\mathbf{y}}^i] = O(n^{1-1/d}/(\varepsilon^2(i-i_0+1)^3))$*.*

Similarly, for each $j \in [m]$ *and each* $i > i_0$*,* $\mathbb{E}[\mathbf{Q}_{j*}^i\tilde{\mathbf{y}}^i] = \mathbf{Q}_{j*}^i\mathbf{y}^i$*, and* $\mathrm{Var}[\mathbf{Q}_{j*}^i\tilde{\mathbf{y}}^i] = O(n^{1-1/d}(i-i_0+1)^3/(2^{i-i_0}\varepsilon^2))$*.*

We're now ready to prove an approximation guarantee.

Theorem 7 (**Utility**). *The expected average squared error of Algorithm 1 is* $O(n^{1-1/d}/\varepsilon^2)$*. With probability at least* $1-\beta$*, the worst-case squared error of Algorithm 1 is at most* $O(n^{1-1/d}\log(n/\beta)/\varepsilon^2)$*.*

Proof. Let $\mathbf{z}^2 = \sum_{i=1}^k \mathbf{Q}^i\mathbf{y}^i$. Note that all $\tilde{\mathbf{y}}^i$ have indepedentent noise. Then, by Lemma 14, for each $j \in [m]$, $\mathbb{E}[\tilde{z}_j^2] = z_j^2$ and $\mathrm{Var}[\tilde{z}_j^2] = O(n^{1-1/d}/\varepsilon^2)$ The expected total squared error of Algorithm 1 is, by linearity of expectation $\sum_j \mathrm{Var}[\tilde{z}_j]$. Since $\tilde{\mathbf{z}}^1$ is independent from $\tilde{\mathbf{z}}^2$, we have $\sum_j \mathrm{Var}[\tilde{z}_j] = \sum_j \mathrm{Var}[\tilde{z}_j^1] + \sum_j \mathrm{Var}[\tilde{z}_j^2]$. By claim 1. in Lemma 13, the first term is the result of a recursive call on input of size at most $n/2$. We can express the expected squared error as a function $E(n)$ recursively as $E(n) = E(n/2) + O(n^{1-1/d}/\varepsilon^2)$ which is easily seen to resolve to $E(n) = O(n^{1-1/d}/\varepsilon^2)$.

The worst-case guarantee can be derived by standard use of tail bounds for sums of Laplace random variables. □

6 Extensions

Algorithms for halfspace range counting can be derived from several other methods, each of which provides weaker noise guarantees and/or less generality.

The partition trees of Chan [6] imply a way to factor the incidence matrix $\mathbf{A}$ of a range space induced by d-dimensional halfspaces into matrices $\mathbf{Q}$ and $\mathbf{D}$ such that $\mathbf{A} = \mathbf{QD}$, each column in $\mathbf{D}$ has at most $O(\log\log n)$ nonzero elements, each row in $\mathbf{Q}$ has at most $O(n^{1-1/d})$ nonzero elements, and $\mathbf{Q}$ and $\mathbf{D}$ both have elements bounded in absolute value by 1. Using Lemma 4, we can add Laplace noise with variance $O(\frac{1}{\varepsilon^2}\log\log n)$ to each element of $\mathbf{Dx}$, preserving $(\varepsilon\sqrt{\ln 1/\delta}, \delta)$ privacy. We can then bound the variance of this mechanism to argue that, with constant probability, the average squared error is $O(\frac{1}{\varepsilon^2}n^{1-1/d}\log\log n)$ and the worst case squared error is $O(\frac{1}{\varepsilon^2}n^{1-1/d}\log n \log\log n)$.

Welzl [31], and Chazelle and Welzl [10] gave an algorithm that, given a set of points P in $\mathbb{R}^d$, computes a spanning path such that any hyperplane intersects the path in at most $O(n^{1-1/d})$ components. Then the intersection of any halfspace with P can be represented as the union of $O(n^{1-1/d})$ disjoint intervals on the spanning path. An algorithm for privately computing interval counting queries, e.g. the algorithm from [5], can be used with the spanning path as input, giving average squared error $O(\frac{1}{\varepsilon^2}n^{1-1/d}\log n)$ and worst case squared error $O(\frac{1}{\varepsilon^2}n^{1-1/d}\log^2 n)$. Interestingly, the spanning path approach generalizes to range spaces whose *dual* shatter function is bounded by a polynomial with exponent d.

There is a well-known connection between combinatorial discrepancy and epsilon approximations (c.f. [26], Chapter 1). Let $(P, \mathcal{R})$ be a range space such that the maximum discrepancy over all restrictions of $\mathcal{R}$ to a size s subset of P is $f(s)$ (this is the same $f(s)$ as in Section 3). Under some reasonable assumptions on the range space, there exists a subset S of P of size s such that range counts on S are close to range counts on P to within an additive $\frac{n}{s}f(s)$. Using this fact, and the discrepancy upper bound for range spaces with shatter function exponent d, we can apply the median mechanism of Roth and Roughgarden [28] with the new analysis in [19] to obtain a squared error upper bound that depends on n as $O(n^{2d/(2d+1)})$. This upper bound is suboptimal; for example, for $d = 2$, it yields an upper bound of $n^{4/5}$ as opposed to the optimal $n^{1/2}$. Nevertheless, this method still gives squared error bounds that grow slower than n for range system with polynomial shatter function. It also extends to the case where the universe is much larger than $\|\mathbf{x}\|_1$. Giving optimal or near optimal error upper bounds in this large universe regime is an interesting open problem.

7 Concluding Remarks

While predicate count queries ($\mathbf{Ax}$) have been studied in differential privacy before, we make one of the first significant progress in understanding the complexity of the problem in terms of the combinatorial properties of $\mathbf{A}$, in particular for halfspace, orthogonal and other range count queries. Our main result is tight upper and lower bounds on approximation of (ϵ, δ) differentially private halfspace count queries. Our approach is via a variation of discrepancy. The main problems we leave open are to get tight bounds for orthogonal counts with (ϵ, δ)-differential privacy and to extend our bounds to the large universe regime.

Acknowledgements

We would like to thank Guy Rothblum, Kobbi Nissim, and Aaron Roth for helpful discussions, and the anonymous reviewers for useful suggestions and corrections.

This material is based upon work supported by the National Science Foundation under Grant No. 0916782.

References

[1] P.K. Agarwal and J. Erickson. Geometric range searching and its relatives. In *Advances in discrete and computational geometry: proceedings of the 1996 AMS-IMS-SIAM joint summer research conference, Discrete and Computational Geometry–Ten Years Later, July 14-18, 1996, Mount Holyoke College*, volume 223, page 1. Amer Mathematical Society, 1999.

[2] Boaz Barak, Kamalika Chaudhuri, Cynthia Dwork, Satyen Kale, Frank McSherry, and Kunal Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '07, pages 273–282, New York, NY, USA, 2007. ACM.

[3] Jozsef Beck. Balanced two-colorings of finite sets in the square. *Combinatorica*, 1(4):327–335, 1981.

[4] Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to non-interactive database privacy. In *Proceedings of the 40th annual ACM symposium on Theory of computing*, STOC '08, pages 609–618, New York, NY, USA, 2008. ACM.

[5] T.H.H. Chan, E. Shi, and D. Song. Private and continual release of statistics. In *ICALP*, 2010.

[6] T.M. Chan. Optimal partition trees. In *Proceedings of the 2010 annual symposium on Computational geometry*, pages 1–10. ACM, 2010.

[7] B. Chazelle. *The discrepancy method*. Cambridge Univ. Press, 2000.

[8] B. Chazelle and A. Lvov. A trace bound for the hereditary discrepancy. *Discrete & Computational Geometry*, 26(2):221–231, 2001.

[9] B. Chazelle, J. Matoušek, and M. Sharir. An elementary approach to lower bounds in geometric discrepancy. *Discrete & Computational Geometry*, 13(1):363–381, 1995.

[10] B. Chazelle and E. Welzl. Quasi-optimal range searching in spaces of finite vc-dimension. *Discrete Comput. Geom.*, 4:467–489, September 1989.

[11] Anindya De. Lower bounds in differential privacy. *CoRR*, abs/1107.2183, 2011.

[12] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '03, pages 202–210, New York, NY, USA, 2003. ACM.

[13] C. Dwork, F. Mcsherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006.

[14] C. Dwork, G. N. Rothblum, and S. Vadhan. Boosting and differential privacy. In *Proc. 51st Annual IEEE Symp. Foundations of Computer Science (FOCS)*, pages 51–60, 2010.

[15] Cynthia Dwork, Frank McSherry, and Kunal Talwar. The price of privacy and the limits of lp decoding. In *STOC*, pages 85–94, 2007.

[16] Cynthia Dwork, Moni Naor, Omer Reingold, Guy N. Rothblum, and Salil Vadhan. On the complexity of differentially private data release: efficient algorithms and hardness results. In *Proceedings of the 41st annual ACM symposium on Theory of computing*, STOC '09, pages 381–390, New York, NY, USA, 2009. ACM.

[17] Cynthia Dwork and Sergey Yekhanin. New efficient attacks on statistical disclosure control mechanisms. In *CRYPTO*, pages 469–480, 2008.

[18] Anupam Gupta, Moritz Hardt, Aaron Roth, and Jonathan Ullman. Privately releasing conjunctions and the statistical query barrier. In *Proceedings of the 43rd annual ACM symposium on Theory of computing*, STOC '11, pages 803–812, New York, NY, USA, 2011. ACM.

[19] Anupam Gupta, Aaron Roth, and Jonathan Ullman. Iterative constructions and private data release. *CoRR*, abs/1107.3731, 2011.

[20] M. Hardt and G. N. Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *Proc. 51st Annual IEEE Symp. Foundations of Computer Science (FOCS)*, pages 61–70, 2010.

[21] Moritz Hardt, Katrina Ligett, and Frank McSherry. A simple and practical algorithm for differentially private data release. *CoRR*, abs/1012.4763, 2010.

[22] Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. In *Proceedings of the 42nd ACM symposium on Theory of computing*, STOC '10, pages 705–714, New York, NY, USA, 2010. ACM.

[23] D. Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995.

[24] Shiva Prasad Kasiviswanathan, Mark Rudelson, Adam Smith, and Jonathan Ullman. The price of privately releasing contingency tables and the spectra of random matrices with correlated rows. In *Proceedings of the 42nd ACM symposium on Theory of computing*, STOC '10, pages 775–784, New York, NY, USA, 2010. ACM.

[25] J. Matoušek. Tight upper bounds for the discrepancy of half-spaces. *Discrete and Computational Geometry*, 13(1):593–601, 1995.

[26] J. Matoušek. *Geometric discrepancy: An illustrated guide*, volume 18. Springer Verlag, 2010.

[27] Aaron Roth. Differential privacy and the fat-shattering dimension of linear queries. In *Proceedings of the 13th international conference on Approximation, and 14 the International conference on Randomization, and combinatorial optimization: algorithms and techniques*, APPROX/RANDOM'10, pages 683–695, Berlin, Heidelberg, 2010. Springer-Verlag.

[28] Aaron Roth and Tim Roughgarden. Interactive privacy via the median mechanism. In *Proceedings of the 42nd ACM symposium on Theory of computing*, STOC '10, pages 765–774, New York, NY, USA, 2010. ACM.

[29] K.F. Roth. On irregularities of distribution. *Mathematika*, 1(02):73–79, 1954.

[30] J. Spencer. Six standard deviations suffice. *Trans. Amer. Math. Soc*, 289, 1985.

[31] Emo Welzl. On spanning trees with low crossing numbers. In *Data Structures and Efficient Algorithms, Final Report on the DFG Special Joint Initiative*, pages 233–249, London, UK, 1992. Springer-Verlag.

[32] Xiaokui Xiao, Guozhang Wang, and J. Gehrke. Differential privacy via wavelet transforms. In *Proc. IEEE 26th Int Data Engineering (ICDE) Conf*, pages 225–236, 2010.

[33] Andrew Chi-Chih Yao and F. Frances Yao. A general approach to d-dimensional geometric queries (extended abstract). In *STOC*, pages 163–168, 1985.